Discovering Statistically Significant Co-location Rules in Datasets with Extended Spatial Objects

Jundong Li[1], Osmar R. Zaïane[1], and Alvaro Osornio-Vargas[2]

[1] Department of Computing Science, University of Alberta, Edmonton, Canada
[2] Department of Paediatrics, University of Alberta, Edmonton, Canada
{jundong1,zaiane,osornio}@ualberta.ca

Abstract. Co-location rule mining is one of the tasks of spatial data mining, which focuses on the detection of sets of spatial features that show spatial associations. Most previous methods are generally based on transaction-free apriori-like algorithms which are dependent on user-defined thresholds and are designed for boolean data points. Due to the absence of a clear notion of transactions, it is nontrivial to use association rule mining techniques to tackle the co-location rule mining problem. To solve these difficulties, a transactionization approach was recently proposed; designed to mine datasets with extended spatial objects. A statistical test is used instead of global thresholds to detect significant co-location rules. One major shortcoming of this work is that it limits the size of antecedent of co-location rules up to three features, therefore, the algorithm is difficult to scale up. In this paper we introduce a new algorithm that fully exploits the property of statistical significance to detect more general co-location rules. We use our algorithm on real datasets with the National Pollutant Release Inventory (NPRI). A classifier is also proposed to help evaluate the discovered co-location rules.

Keywords: Co-location Rules, Statistically Significant, Classifier.

1 Introduction

Co-location mining, one of the canonical tasks of spatial data mining, has received increasing attention in recent years. It tries to find a set of spatial features that are frequently co-located together, i.e. in a geographic proximity. A motivating application example is the detection of possible co-location rules between chemical pollutants and cancer cases with children. Previous work [13,15,14,12] are mainly based on transaction-free algorithms with an apriori-like framework. A prevalence measure threshold is required in the property of anti-monotonicity for effective pruning, the strength of co-location rules are determined afterwards with a prevalence measure threshold. However, the support-confidence framework fails to capture the statistical dependency between spatial features. On one hand, the antecedent and consequent spatial features may be independent of each other. On the other hand, some other strong dependent co-location rules may be ignored due to a prevalence measure value. In the worst case, all detected

L. Bellatreche and M.K. Mohania (Eds.): DaWaK 2014, LNCS 8646, pp. 124–135, 2014.

co-location rules can be spurious, and strong co-location rules are totally missing. Another limitation of transaction-free apriori-like co-location mining algorithms is that they use only one distance threshold to determine the neighbourhood relationship. However, in real applications, a proper distance threshold is hard to determine. Meanwhile, with only one distance threshold, the neighbourhood relationship among spatial features cannot be fully exploited. For instance, the contaminated area around a chemical facility is affected by the amount of chemical pollutants the facility emits. It is apparently that the more amount of chemical pollutants it emits, the more neighbourhood relationships it should capture.

To solve the previous mentioned limitations of transaction-free apriori-like co-location mining algorithms, Adilmagambetov et al. [2] proposed a new transaction based framework to discover co-location rules in datasets with extended spatial objects. Buffers are built around each spatial object, the buffer zone could be the same for all spatial objects or it might be affected by some other spatial or non-spatial features, like the amount of chemical pollutants the facility emits, wind direction in this region, etc. Then, grids are imposed over the geographic space; each grid point intersecting with a set of spatial objects could be seen as a transaction. As mentioned above, the usage of support-confidence framework may result in the discovery of incorrect co-location rules and omission of strong co-location rules. Therefore, to find statistically significant co-location rules, a statistical test method is used instead of global thresholds. However, the statistical significance is not a monotonic property and it cannot be used to prune insignificant co-location rules as apriori-like algorithms. Thus in their work, they limit the size of the antecedent of a rule up to three features and test each possible candidate co-location rule to see if it passes the statistical test. The algorithm cannot scale up well for co-location rules with more than three spatial features in the antecedent, and therefore limits its use.

In this paper, we investigate how to exploit the property of statistical significance to scale it up to detect more general co-location rules. We propose a new algorithm: Co-location Mining Constrained StatApriori (CMCStatApriori) which is able to detect statistically significant co-location rules without any limitation on the rule size. CMCStatApriori is based on the work of StatApriori [8,10]. It uses the z-score to search for statistically significant co-location rules with a fixed consequent spatial feature. The results of co-location rules are hard to evaluate even for domain experts, therefore, we also propose to use a classifier to help evaluate the results of co-location rules.

The remainder of the paper is organized as follows. The overview of related work is given in Section 2. The algorithm framework is described in Section 3. Section 4 describes the experimental results and the evaluation of the results. Section 5 concludes the paper.

2 Related Work

In this section, we review some related work on co-location mining from two perspectives: the support-confidence framework and the statistical test framework.

2.1 Support-confidence Based Co-location Mining

Shekhar and Huang [13] proposed a co-location pattern mining framework which is based on neighbourhood relations and the concept of participation index. The basic concept of this method is similar to the concept of association rule mining. As an input, the framework takes a set of spatial features and a set of instances, where each instance is a vector that contains information on the instance ID, the feature type of the instance, and the location of the instance. As an output, the method returns a set of co-location rules. A co-location rule is of the form of $C_1 \rightarrow C_2(PI, cp)$, where C_1 and C_2 are a set of spatial features, PI is the prevalence measure (participation index), and cp is the conditional probability. A co-location pattern is considered as prevalent, or interesting, if for each feature of the pattern at least $PI\%$ instances of that feature form a clique with the instances of all other features of the pattern according to the neighbourhood relationship. Similar to association rule mining, only frequent $(k-1)$-patterns are used for the k-candidate generation process. Yoo and Shekhar [15] proposed a join-less algorithm which decreases the computation time of constructing neighbourhood relationship. The main idea is to find star neighbourhoods instead of calculating pairwise distances between all instances in the dataset. Huang et al. [12] continued their previous work by introducing an algorithm that finds co-location patterns with rare features. Instead of the participation index threshold, the authors proposed to use the maximal participation ratio threshold. Briefly, a co-location pattern is considered prevalent if $maxPR\%$ instances of at least one of the features in the pattern are co-located with instances of all other features, where $maxPR$ is the maximal participation ratio. Xiong et al. [14] introduced a framework for detecting patterns in datasets with extended spatial objects. Extended spatial objects are objects that are not limited to spatial points but also include lines and polygons. In the proposed buffer-based model, the candidate patterns are pruned by the coverage ratio threshold. In other words, if the area covered by the features of a candidate pattern is greater than a predefined threshold, this pattern is considered as prevalent or interesting.

2.2 Statistical Test Co-location Mining

The approaches mentioned above use thresholds on interestingness measures, which result in meaningless patterns when a low threshold is used, and a high threshold may prune interesting but rare patterns. Instead of a threshold-based approach, Barua and Sander [4] used the statistical test to mine statistically significant co-location patterns. The participation index of a pattern in the observed data is calculated as previous studies. Then for each co-location pattern the authors compute the probability p of seeing the same or greater value of prevalence measure under a null hypothesis model. The co-location pattern is considered statistically significant if $p \leq \alpha$, where α is a level of significance. Adimagambetov et al. [2] proposed a transactionization framework to find significant co-location rules on extended spatial objects. Spatial instances are transformed into transactions by buffers and grids and the expected support is used as the interesting measure. The statistical test method they used is similar to [4].

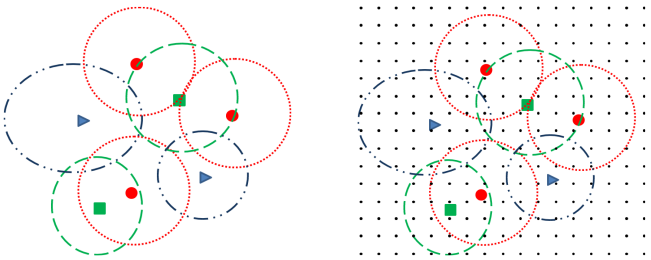

Fig. 1. Transactionization step: (a) An example of spatial dataset with point feature instances and their buffers; (b) Grids imposed over the space.

3 Algorithm Framework

3.1 Problem Definition

The objective is to discover statistically significant co-location rules between a set of antecedent spatial features and one single fixed consequent spatial feature. A real world application of this task is to detect co-location rules between chemical pollutants (antecedent) and cancer cases or other morbidities (consequent). Since we do not intend to find the causality relationships, the goal is to identify potential interesting co-location associations in order to state hypotheses for further study.

The task consists of three steps. In the initialization step, a buffer is built around each spatial object, and it defines the area affected by that object; for example, the buffer zone around an emission point shows the area polluted by a released chemical pollutant. The buffer shape is defined as circle, but it may also be affected by some other factors like wind direction. Considering the factor of wind direction, the circular buffer is transformed to elliptical. Fig. 1(a) displays an example of spatial dataset with buffers of various sizes (circular and elliptical) that are formed around spatial point objects. In the transactionization step, the transaction dataset is formed by imposing grids over all the buffer zones, as shown in Fig. 1(b). Then a transaction is defined as a set of spatial features corresponding to these objects [2]. After getting the derived transaction dataset T from the spatial dataset, we intend to detect statistically significant co-location rules in the next step.

3.2 Co-location Mining Constrained StatApriori

In this subsection, we introduce the proposed Co-location Mining Constrained StatApriori (CMCStatApriori) algorithm which is able to detect statistically significant co-location rules without any rule length limitation.

CMCStatApriori is a variation of StatApriori [8,10]; the main difference is that CMCStatApriori can efficiently detect more specific co-location rules, rules

with one fixed consequent feature. Moreover, the non-redundancy definition in StatApriori is not very practical, it is much more restrictive than the normal definition. Therefore, in CMCStatApriori, we do not intend to target for non-redundant significant co-location rules.

For the co-location rule $X \rightarrow A$ ($F = \{f_1, ..., f_m\}$ is the set of spatial features and $X \subsetneq F$, $A \in F$), the significance of dependency between X and A is compared with the null hypothesis in which X and A are independent. The statistical significance of the dependency is measured by the p-value, i.e. the probability of observing higher or equal frequency of X and A under null hypothesis. Suppose in the derived transaction dataset T, each transaction can be viewed as an independent Bernoulli trial with two possible results, that $P(XA) = 1$ or $P(XA) = 0$. Thus, the statistical significance of the frequency of XA follows the binomial distribution and the p-value can be formulated as:

$$p = \sum_{i=\sigma(XA)}^{\sigma(A)} \binom{n}{i} (P(X)P(A))^i (1 - P(X)P(A))^{n-i} \tag{1}$$

where $\sigma(XA)$ is the observed frequency of XA, and n is the total number of transactions in T.

The p-value is not a monotonic property, but z-score provides an upper bound for the binomial distribution:

$$z(X \rightarrow A) = \frac{\sigma(XA) - \mu}{s} = \frac{\sqrt{nP(XA)(\gamma(XA) - 1)}}{\sqrt{\gamma(XA) - P(XA)}} \tag{2}$$

where $\mu = nP(X)P(A)$, $s = \sqrt{nP(X)P(A)(1 - P(X)P(A))}$ are the mean and standard deviation of the binomial distribution, respectively. $\gamma(XA) = \frac{P(XA)}{P(X)P(A)}$ is the lift for the co-location rule $X \rightarrow A$. It measures the strength of the dependency between X and A such that $\gamma(X \rightarrow A) > 1$ if X and A show a positive correlation. It is easy to notice that the z-score is a monotonically increasing function with the support and lift of XA: $\sigma(XA)$ and $\gamma(XA)$, therefore, it can be denoted as $z(X \rightarrow A) = f(\sigma(XA), \gamma(XA))$.

Therefore, following StatApriori [8,10], the search problem can be reformulated as searching for all statistically significant co-location rules in the form of $X \rightarrow A$ with the following requirements (the set of statistically significant co-location rules is denoted as P):

Definition 1. *Statistically significant co-location rules*

1. $X \rightarrow A$ *expresses a positive correlation, i.e.* $\gamma(X \rightarrow A) > 1$
2. *for all* $(Y \rightarrow A) \notin P, z(X \rightarrow A) > z(Y \rightarrow A)$
3. $z(X \rightarrow A) \geq z_{min}$

With this definition, the property "potentially significant" (PS) is defined as follows. It is a necessary condition to construct the set of statistically significant co-location rules.

Definition 2. *Let A be the fixed consequent feature, z_{min} is an user-defined threshold for the z-score, and upperbound(f) be an upper bound for the function f. The co-location rule $X \rightarrow A$ is defined as potentially significant, i.e. $PS(X) = 1$, iff $upperbound(z(X \rightarrow A)) \geq z_{min}$. Otherwise, the co-location rule is not considered as statistically significant.*

The property of PS displays a monotonic property in some specific situations:

Theorem 1. *Let A be the fixed consequent feature and $PS(X) = 1$, then for all $Y \subseteq X$ and $min(XA) = min(YA)$ we can get $PS(Y) = 1$, where $min(XA)$ denotes the feature with the minimum support in XA.*

The proof of Theorem 1 is straightforward, first we can see that:

$$\gamma(YA) = \frac{P(YA)}{P(Y)P(A)} \leq \frac{1}{P(Y)} \leq \frac{1}{P(min(YA))} \tag{3}$$

where $min(YA)$ denotes the feature with the smallest support among YA, the upper bound of the co-location rule $Y \rightarrow A$ now is:

$$upperbound(z(Y \rightarrow A)) = f(P(YA), \frac{1}{P(min(YA))}) \tag{4}$$

then we have:

$$\begin{aligned} &upperbound(z(X \rightarrow A)) = f(P(XA), \frac{1}{P(min(XA)}) \leq \\ &f(P(YA), \frac{1}{P(min(YA)}) = upperbound(z(Y \rightarrow A)) \end{aligned} \tag{5}$$

for all $Y \subseteq X$ such that $min(XA) = min(YA)$. We can see that the monotonic property is kept only when the minimum feature (the feature with the minimal support) in XA and YA are the same.

With the monotonic property of PS, we can derive the algorithm that discovers the potential significant co-location rules in the same way as the general Apriori-like algorithms do, alternating between the candidate generation and candidate pruning. First, the set of antecedent features are arranged in an ascending order by their frequencies. Let the renamed features be $\{f'_1, f'_2, ..., f'_{m-1}\}$, where $P(f'_1) \leq P(f'_2) \leq ... \leq P(f'_{m-1})$. The candidate generation process is the same as that in Apriori [3], for the l-set $S_l = \{f'_{a_1}, ..., f'_{a_l}\}$ $(a_1 < a_2 < ... < a_l)$, we can generate $(l+1)$-sets $S_l \cup \{f'_{a_j}\}$, where $a_j > a_l$. After the generation of the $(l+1)$-sets $S_l \cup \{f'_{a_j}\}$, we need to check if all of its l-set "regular" parents (the parents with the same minimum support feature when combined with A as $S_l \cup \{f'_{a_j}\} \cup A$) can indicate PS co-location rules. If all of its regular parents can indicate PS co-location rules, then $S_l \cup \{f'_{a_j}\}$ is added to the candidate set for the pruning process, otherwise, $S_l \cup \{f'_{a_j}\}$ can be pruned directly. In the pruning process, the PS co-location rule $X \rightarrow A$ is kept if it meets the z_{min} threshold, otherwise, it is removed.

Algorithm 1. CMCStatApriori Algorithm

Require: Set of antecedent features $F\backslash A$, the consequent feature A, derived transaction dataset T, the threshold z_{min} for the z-score
Ensure: Set of potentially significant co-location rules P
1: $P_1 = \{f_i \in F\backslash A | PS(f_i) = 1\}$
2: $l = 1$
3: **while** $(P_l \neq \varnothing)$ **do**
4: $C_{l+1} = GenCands(P_l, A)$
5: $P_{l+1} = PrunCands(C_{l+1}, z_{min}, A)$
6: $l = l + 1$
7: **end while**
8: $P = \cup_i P_i$
9: **return** P

Algorithm 2. Algorithm GenCands

Require: Potentially significant l-sets P_l, the consequent feature A.
Ensure: $(l+1)$-candidates C_{l+1}.
1: $C_{l+1} = \emptyset$
2: **for** all $Q_i, Q_j \in P_l$ such that $|Q_i \cap Q_j| = l - 1$ **do**
3: **if** $\forall Z \subseteq Q_i \cup Q_j$ such that $|Z| = l$ and $min(ZA) = min((Q_i \cup Q_j)A)$ and $Z \subseteq P_l$ **then**
4: $C_{l+1}.add(Q_i \cup Q_j)$
5: **end if**
6: **end for**
7: **return** C_{l+1}

Algorithm 3. Algorithm PruneCands

Require: l-candidates C_l, threshold z_{min}, the consequent feature A.
Ensure: Potentially significant l-sets P_l
1: $P_l = \varnothing$
2: **for** all $Q_i \in C_l$ **do**
3: calculate $P(Q_iA)$ and the upperbound of lift $\frac{1}{P(min(Q_iA))}$
4: **if** $f(P(Q_iA), \frac{1}{P(min(Q_i,A))}) \geq z_{min}$ **then**
5: $P_l.add(Q_i)$
6: **end if**
7: **end for**
8: **return** P_l

A problem of StatApriori is that for each potentially significant set C, only the best rule is derived from C. For example, if $C\backslash A \rightarrow A$ is the best rule, where $A \in C$ and the "best" indicates that the rule has the highest z-score, then no other rules in the form of $C\backslash B \rightarrow B (B \neq A)$ is output. However, in our CMCStatApriori algorithm, this kind of problem does not exist, because the PS property is for the co-location rule and the consequent feature is fixed. The detailed pseudo code of CMCStatApriori is illustrated in Algorithms 1, 2 and 3.

4 Experiments

4.1 Datasets

We conduct our experiments on two real datasets which contain pollutant emissions and information about cancer cases for children in the provinces of Alberta and Manitoba, Canada. The sources of the data are the National Pollutant Release Inventory (NPRI) [5] and the provincial cancer registries. The information on pollutants is taken for the period between 2002 and 2007 and contains the type of a chemical, location of release, and average amount of release per year. In order to get reliable results, the chemical pollutants that had been emitted from less than three facilities are excluded from the dataset. There are 47 different chemical pollutants and 1,422 chemical pollutant emission points in Alberta; 26 different chemical pollutants and 545 chemical pollutant emission points in Manitoba, several chemical pollutants might be released from the same location. The number of cancer cases are 1,254 and 520 in Alberta and Manitoba, respectively. In order to make the model more accurate, the wind speed and direction are also taken into account in these two provinces. The interpolation of wind information between wind stations is used. In Alberta, the data of 18 stations are from Environmental Canada [6] and 156 stations are from ArgoClimatic Information Service (ACIS) [1]. In Manitoba, the data of all 20 stations are all from Environment Canada [6]. We obtain the wind direction and speed in the locations of chemical facilities by making interpolations in the ArcGIS tool [7].

4.2 Experimental Settings

We are interested in co-location rules of the form of $Pol \rightarrow Cancer$, where Pol is a set of pollutant features and $Cancer$ is a cancer feature. Three different methods are compared: the co-location mining algorithm by Adilmagambetov et al. in [2] (denoted as CM), co-location mining algorithm with Kingfisher [9,11] (denoted as CMKingfisher) and the proposed CMCStatApriori method. In all of these three methods, the distance between grid points is 1km.

CM. CM needs a number of simulations to detect significant co-location rules, the number of simulations for the statistical test is set to be 99 and the level of significance α is set to be 0.05. The size of antecedent features of a candidate rule is up to three. The randomized datasets (simulations) that are used in the statistical test are generated according to the distributions of chemical pollutant emitting facilities and cancer cases. Chemical pollutant emitting facilities are not randomly distributed, and are usually located close to regions with high population density, thus, CM does not randomize the pollutant facilities all over the region, instead, it keeps locations of facilities and randomize the pollutants within these regions. For the cancer cases, most of them are located within dense "urban" regions and the rest are in "rural" regions. Therefore, the cancer cases are randomized according to the population ratio of "urban" regions to "rural" regions. In each simulation of CM, both pollutant chemicals and cancer cases are randomized.

CMKingfisher. Kingfisher [9,11] is developed to discover positive and negative dependency rules between a set of antecedent features and a single consequent feature. The algorithm is based on a branch and bound strategy to search for the best, non-redundant dependency top-K rules. Kingfisher is able to detect statistically significant positive and negative rules with any possible consequent. But we are only interested in the positive rules whose consequent is "Cancer", therefore, after getting the derived transaction dataset T, we apply Kingfisher algorithm to get the complete set of co-location rules and extract the subset of co-location rules that we are interested in. The significance level α is 0.05.

CMCStatApriori. The CMCStatApriori is the algorithm proposed in this paper. Unlike CM and CMKingfisher which use the p-value as a significance level, CMStatApriori uses the z-score which provides an upper bound for the p-value. In the experiment, the threshold of z-score is set to be 150 in the Alberta dataset. This threshold of 150 is too high in the Manitoba dataset and no co-location rules are output. Therefore, we set a lower z-score threshold of 40. Indeed, the lower the z-score threshold, the more co-location rules is generated. The parameter setting of z-score threshold of CMCStatApriori is discussed in the last subsection.

4.3 Experimental Results

Both CMKingfisher and CMCStatApriori are able to detect more general co-location rules (without limitation of size of antecedent features). However, to have a fair comparison with CM, we only list the co-location rules with up to three antecedent features. The number of rules detected by these three methods and the number of rules overlaps with CM by CMKingfisher as well as CMCStat-Apriori are listed in Table 1. It can be observed that in the dataset of Alberta, both of CMKingfisher and CMCStatApriori have a small overlap with CM rules. The situation is slightly different in the dataset of Manitoba, around 80% and 30% of detected rules by CMKingfisher and CMCStatApriori also appear in CM.

Table 1. Number of co-location generated by different methods

	Alberta		Manitoba	
	#rules	# rules in CM	rules	# rules in CM
CM	273	–	170	–
CMKingfisher	108	7	23	19
CMCStatApriori	571	5	60	16

4.4 Evaluation

Environmental pollutants are suspected to be one of the causes of cancer in children. However, there are other factors that could lead to this disease. Therefore, it is a difficult task to evaluate the detected co-location rules even for domain experts. To assist in evaluating the discovered co-location rules, we propose to use

a classifier with the co-location rules as a predictive model. The results by different methods are carefully and painstakingly evaluated manually by experts in our multidisciplinary team. However, the systematic evaluation by classification provides an estimation of the best quality co-location rule set.

We consider co-location rules generated by either method as a classifier. To evaluate the discovered co-location rules, we randomly sample some grid points on the geographic space. The randomly sampled grid point has to intersect with at least one pollutant feature; it either intersects with cancer or not. For the type of grid point $(Pol_{grid}, Cancer)$ intersects with both pollutant(s) and cancer, if we can find at least one co-location rule $Pol \rightarrow Cancer$ in the classifier that correctly matches it, i.e. $Pol \subseteq Pol_{grid}$, the grid point is indicated as correctly classified. For the other type of grid point $(Pol_{grid}, \neg Cancer)$ intersects with pollutant(s) but not cancer, if there does not exist any co-location rules $Pol \rightarrow Cancer$ that match it, i.e. $Pol \nsubseteq Pol_{grid}$, the grid point is also indicated as correctly classified. Otherwise, the grid points are considered as misclassified. The ratio of correctly classified grid points to the total number of sampled grid points is output as the classification accuracy. Fig. 2 shows a toy example of the evaluation process. In the datasets of Alberta and Manitoba, we randomly sample 1000 grid points each time, repeats 100 times, and calculate the average classification accuracy for the previously mentioned three methods. Table 2 presents the evaluation results, along with the classification accuracy (ACC), the specificity (SPE) and sensitivity (SEN) are also listed. As can be observed from the classification accuracy, CMCStatApriori is better than CM and CMKingfisher. The classification accuracy is much higher in Alberta compared with Manitoba. One possible explanation is that the co-location association between chemical pollutants and children cancer cases is stronger in Alberta. Both the number of rules and the accuracy is very low in Manitoba, therefore, it is possible that chemical pollutants and children cancer cases are more likely to be independent in Manitoba. We can also notice that the specificity is much higher than the sensitivity in both datasets. High specificity means that grid points without cancer are seldom misclassified; on the other hand, low sensitivity indicates that grid points with cancer are mostly misclassified. This phenomenon may imply that the co-location associations between chemical pollutants and children cancer cases is weak. However, these assumptions still need to be carefully scrutinized.

Table 2. Evaluation of different methods using Accuracy, Specificity and Sensitivity

	Alberta			Manitoba		
	ACC	SPE	SEN	ACC	SPE	SEN
CM	83.9 ± 3.3	97.6 ± 1.6	11.4 ± 8.1	22.0 ± 4.3	55.8 ± 11.2	$\mathbf{13.4 \pm 3.7}$
CMKingfisher	69.2 ± 4.1	77.4 ± 4.1	$\mathbf{28.6 \pm 11.4}$	26.6 ± 4.6	$\mathbf{96.4 \pm 3.6}$	8.7 ± 3.0
CMCStatApriori	$\mathbf{84.7 \pm 3.4}$	$\mathbf{99.6 \pm 0.7}$	6.6 ± 6.4	$\mathbf{27.4 \pm 4.1}$	83.4 ± 7.7	12.2 ± 3.4

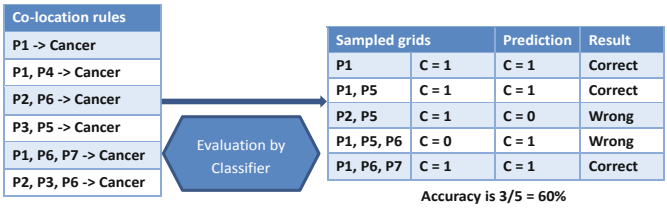

Fig. 2. Toy example of the classification evaluation

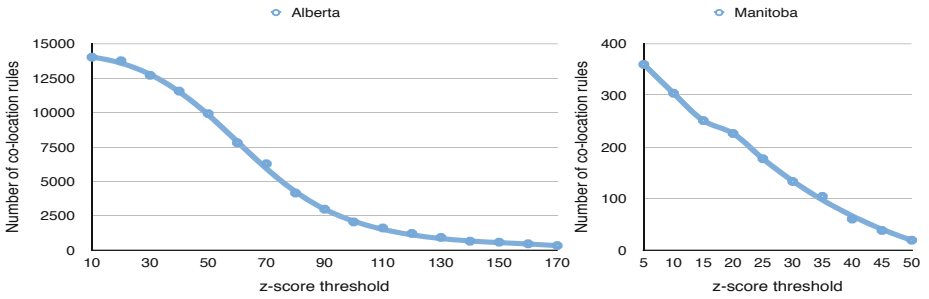

Fig. 3. Number of co-location rules on Alberta and Manitoba dataset

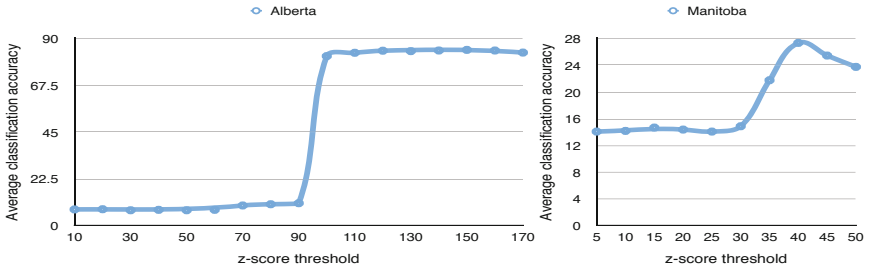

Fig. 4. Average classification accuracy of sampled grids

The only parameter in CMCStatApriori is the z_{min}. In this subsection, we discuss the effect of the parameter z_{min}. As shown in Fig. 3, the number of discovered co-location rules drops when we increase z_{min}. We were not able to find any statistically significant co-location rules when $z_{min} > 170$ in Alberta and when $z_{min} > 50$ in Manitoba. In Fig. 4, the average classification accuracy of the sampled grid points is presented. The classification performance is poor when the z-score threshold is set to be low. Besides, there exists a turning point ($z_{min} = 100$ in Alberta, $z_{min} = 30$ in Manitoba) where the accuracy improves dramatically. In the Alberta dataset, there is not much difference when z_{min} varies from 110 to 170, while in the Manitoba dataset, the performance is best when z_{min} is set to be 40.

5 Conclusion

In this paper, we propose a novel co-location mining algorithm to detect statistically significant co-location rules in datasets with extended spatial objects. By exploiting the property of statistical significance, we do not have to limit the number of antecedent features up to three in co-location rules which is a major shortcoming of previous work. Therefore, more general co-location rules can be generated and the algorithm is able to scale up well. In addition, we propose to use a classifier to help the evaluation of discovered co-location rules.

References

1. AgroClimatic Information Service (ACIS). Live alberta weather station data, http://www.agric.gov.ab.ca/app116/stationview.jsp
2. Adilmagambetov, A., Zaiane, O.R., Osornio-Vargas, A.: Discovering co-location patterns in datasets with extended spatial objects. In: Bellatreche, L., Mohania, M.K. (eds.) DaWaK 2013. LNCS, vol. 8057, pp. 84–96. Springer, Heidelberg (2013)
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: VLDB 1994, pp. 487–499 (1994)
4. Barua, S., Sander, J.: SSCP: Mining statistically significant co-location patterns. In: Pfoser, D., Tao, Y., Mouratidis, K., Nascimento, M.A., Mokbel, M., Shekhar, S., Huang, Y. (eds.) SSTD 2011. LNCS, vol. 6849, pp. 2–20. Springer, Heidelberg (2011)
5. Environment Canada. National Pollutant Release Inventory. Tracking Pollution in Canada, http://www.ec.gc.ca/inrp-npri/
6. Environment Canada. National Climate Data and Information. Canadian climate normals or averages 1971-2000, http://climate.weatheroffice.gc.ca/climate_normals/index_e.html
7. ESRI. ArcGIS Desktop: Release 10 (2011)
8. Hämäläinen, W., Nykanen, M.: Efficient discovery of statistically significant association rules. In: ICDM, pp. 203–212 (2008)
9. Hämäläinen, W.: Efficient discovery of the top-k optimal dependency rules with fisher's exact test of significance. In: ICDM, pp. 196–205 (2010)
10. Hämäläinen, W.: Statapriori: an efficient algorithm for searching statistically significant association rules. KAIS 23(3), 373–399 (2010)
11. Hämäläinen, W.: Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. KAIS 32(2), 383–414 (2012)
12. Huang, Y., Pei, J., Xiong, H.: Mining co-location patterns with rare events from spatial data sets. Geoinformatica 10(3), 239–260 (2006)
13. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: A summary of results. In: Jensen, C.S., Schneider, M., Seeger, B., Tsotras, V.J. (eds.) SSTD 2001. LNCS, vol. 2121, pp. 236–256. Springer, Heidelberg (2001)
14. Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., Yoo, J.S.: A framework for discovering co-location patterns in data sets with extended spatial objects. In: SDM (2004)
15. Yoo, J.S., Shekhar, S.: A joinless approach for mining spatial co-location patterns. TKDE 18(10), 1323–1337 (2006)